Online Prediction of User Enjoyment in Human-Robot Dialogue with LLMs

Ruben Janssens
IDLab-AIRO, Ghent University – imec
Ghent, Belgium
ruben.janssens@ugent.be

André Pereira
KTH Royal Institute of Technology
Stockholm, Sweden
atap@kth.se

Gabriel Skantze
KTH Royal Institute of Technology
Stockholm, Sweden
skantze@kth.se

Bahar Irfan
KTH Royal Institute of Technology
Stockholm, Sweden
birfan@kth.se

Tony Belpaeme
IDLab-AIRO, Ghent University – imec
Ghent, Belgium
tony.belpaeme@ugent.be

***Abstract*—Large Language Models (LLMs) allow social robots to engage in unconstrained open-domain dialogue, but often make mistakes when employed in real-world interactions, requiring adaptation of LLMs to specific conversational contexts. However, LLM adaptation techniques require a feedback signal, ideally for multiple alternative utterances. At the same time, human-robot dialogue data is scarce and research often relies on external annotators. A tool for automatic prediction of user enjoyment in human-robot dialogue is therefore needed. We investigate the possibility of predicting user enjoyment turn-by-turn using an LLM, giving it a proposed robot utterance within the dialogue context, but without access to user response. We compare this performance to the system's enjoyment ratings when user responses are available and to assessments by expert human annotators, in addition to self-reported user perceptions. We evaluate the proposed LLM predictor in a human-robot interaction (HRI) dataset with conversation transcripts of 25 older adults' 7-minute dialogues with a companion robot. Our results show that an LLM is capable of predicting user enjoyment, without loss of performance despite the lack of user response and even achieving performance similar to that of human expert annotators. Furthermore, results show that the system surpasses expert annotators in its correlation with the user's self-reported perceptions of the conversation. This work presents a tool to remove the reliance on external annotators for enjoyment evaluation and paves the way toward real-time adaptation in human-robot dialogue.**

***Index Terms*—human-robot interaction; open-domain dialogue; user enjoyment; prediction; large language model**

I. Introduction

Open-domain dialogue is a long-held dream and necessity in human-robot interaction. Wherever socially interactive robots are employed, such as companion robots in elderly care [1] or tutors for children in classrooms [2], they need to be able to engage in conversations with their users, as this is the most natural communication interface for humans. Large Language Models (LLMs) have recently provided the technology that allows robots to react to anything a user could say, without needing a script. However, they are known to often fail in ways that disrupt the conversational flow when integrated in social robots for conversational interactions, such as repeatedly asking the same questions, giving replies that do not invite further conversation, or hallucinating, all of which can negatively affect user experiences, such as enjoyment during the conversation [3].

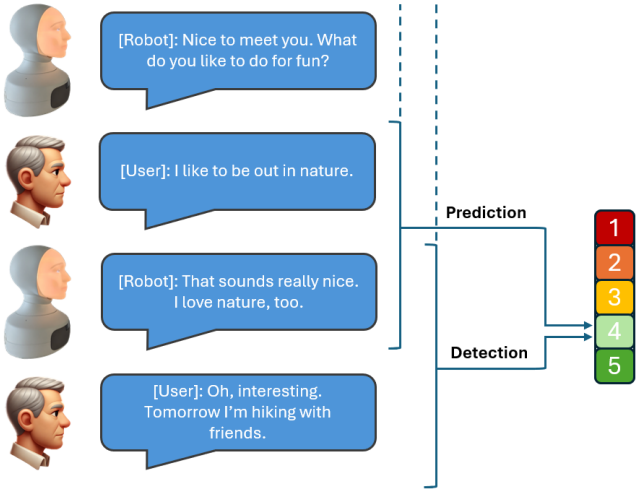

Fig. 1. We compare an LLM's performance at rating user enjoyment in two conditions: *detecting* user enjoyment from an exchange that contains the user's response to a robot utterance, or *predicting* what the user enjoyment will be after a given robot utterance. The diagram shows what the input is to the system in those two conditions, with the dashed line indicating the dialogue context that is also provided.

In order to successfully drive social robot dialogue, LLMs need this other fundamental human capacity: adapting to the specific user and situation they are employed in—also called alignment [4]. Real-time adaptation methods could manage dialogue and prevent a conversation from going awry by choosing the most appropriate potential response at any time [5], while offline adaptation methods such as Reinforcement

This research received funding from the Flemish Government (AI Research Program), the Research Foundation Flanders - FWO (grant V449824N), and KTH Digital Futures (Sweden).

Learning from Human Feedback (RLHF) [6], [7] could help to specialise models for use in specific application contexts, or could be a step towards lifelong learning or personal robots [8].

However, any such adaptation techniques require a feedback signal: an indication of how much the user will enjoy the dialogue, and which potential utterance would be the best. This clearly outlines a bottleneck that stands in the way of applying LLM adaptation techniques to social robotics: we need to be able to automatically predict user enjoyment on a future robot utterance, without needing a human annotator. In this work, we show that this is possible. We build a system that predicts turn-by-turn ratings for user enjoyment in human-robot dialogues. User enjoyment is rated on a 5-point scale that was developed and validated by experts in comparison to self-reported user perceptions in [9], named the Human-Robot Interaction Conversational User Enjoyment Scale (HRI CUES). The system is evaluated on a corpus of 25 open-domain conversations between older adults and a Furhat social robot whose dialogue was generated by an LLM, totalling 590 dialogue exchanges, each of which was rated by three expert annotators, as also released by [9] alongside the rating scale.

We aim to answer the following research question: *Can turn-by-turn user enjoyment be predicted without the availability of the user response to a robot utterance?* We compare the performance of a system that does not have access to the user's response (the enjoyment "prediction" system) to one that does have this information ("detecting" enjoyment instead), as shown in Figure 1. We hold these systems' performances to the standard of expert external human annotators, but also look at correlation with the user's own perceptions, based on user-reported questionnaire results on enjoyment, which had been completed after the interaction. We publicly release our code[1], aiming to fast-forward research into the adaptation of LLMs for human-robot dialogue.

II. Related Work

In general, there is a lack of systems that autonomously evaluate dialogue quality [10] or user enjoyment in dialogue [11], especially in social robotics. Despite the large interest in LLMs as chatbots, evaluation of conversation quality is usually done through pairwise comparisons, performed by humans or LLMs [12]. While prior research has explored automatically detecting user enjoyment or conversational quality in human-robot interaction, to the best of our knowledge, none have explored predicting it for utterances without a user response.

Paetzel-Prüsmann et al. [13] built a system to automatically evaluate conversation quality for multi-party, multi-session social dialogues with robots. Their analysis happens on a whole-dialogue level instead of turn-by-turn and is designed for conversational agents using dialogue trees and intent recognition algorithms, not for an open-domain LLM-driven dialogue system.

[1] github.com/rubenjanss/enjoyment-prediction-public

Specifically towards enjoyment in conversations, instead of general conversation quality, research has mostly focussed on detecting emotions in user responses using supervised models [14], [15] or LLMs [16], [17]. Recent research has shown that LLMs are also able to reason about a person's emotions from a third-party perspective, reaching the same as average human performance [18].

Pereira et al. [11] investigated automatically detecting the user enjoyment of a conversation turn-by-turn. Using exchanges that contain the user's response to the robot utterance, they found that LLMs can achieve similar performance to human annotators on this task. However, as this technique requires the user's response to detect enjoyment, it is limited in its application to adapting LLMs: it can only be used reactively to adjust a conversation when low enjoyment is detected, while a predictive system could prevent low enjoyment or be used to adapt an entire LLM to generate more enjoyable dialogue regardless of one particular interaction. Here, we build further on their work: we replicate their detection system with a newer LLM, adapt the system to predict enjoyment without the user response, and compare the two on the same dataset they used.

III. Methodology

We first describe the dataset we use to evaluate our approach and then describe the evaluation metrics that are used. Then, we describe our system and each of the input components we test in an ablation study.

A. Dataset

We evaluate our system on the publicly available HRI CUES dataset [3] of 25 human-robot conversations, each between an older adult and a Furhat social robot. Each conversation lasts for approximately seven minutes and consists of open-domain dialogue: the robot starts with a friendly greeting and the users talk about anything they wanted. The conversations are in Swedish (the native language of the users), and the robot's dialogue is generated by OpenAI's GPT-3.5. The dataset contains transcribed conversation scripts that are split up into exchanges (robot utterance and the subsequent user utterance) for a total of 590 exchanges. Each dialogue consists of 12 to 29 exchanges.

The dataset contains ratings from three expert annotators for each exchange in the dialogue on the five-point HRI CUES scale [9]:

1) Very low enjoyment: discomfort and/or frustration.
2) Low enjoyment: boredom or interaction failure.
3) Neutral enjoyment: politely keeping up the interaction.
4) High enjoyment: smooth and effortless interaction.
5) Very high enjoyment: immersion in the conversation and/or deeper connection with the robot.

This scale was validated on the same dataset we use here, with the three annotators reaching moderate to good agreement [9]. Note that the annotator ratings were based on the user's enjoyment, deriving from their verbal and non-verbal reactions, but not on the robot's speech. As such, even if the robot made a mistake, such as a repeated question or an interruption, this

could still result in a high enjoyment rating if, for example, the user found this to be funny.

Besides turn-by-turn ratings, annotators also rated each whole conversation for enjoyment on a five-point scale. The users themselves were also asked to rate their whole conversation on four separate five-point Likert scale items: *satisfaction*, *fun*, *interestingness*, and *strangeness*.

The four items showed high reliability (Cronbach's $\alpha = 0.84$), but no single external annotator's whole-conversation rating significantly correlated with any of these four items nor the average of these items. Only the average whole-conversation rating of all three annotators showed a significant moderate ($r = -0.42$) Spearman correlation with the users' self-reported perception of strangeness [9].

B. Evaluation

We compare the system's performance in the prediction and detection conditions by evaluating its turn-by-turn enjoyment ratings against those of the human expert annotators. We also report correlations with the whole-conversation ratings reported by the users themselves.

For the turn-by-turn enjoyment ratings, the ratings of the third annotator are considered as ground truth ratings. This annotator's ratings followed the most balanced distribution across the scale out of all three annotators. As an indication of human performance and agreement on this task, we also report the "performance" of the other two annotators when compared with the third annotator's ratings.

The ground truth rating for the prediction and detection conditions are the same: in the prediction condition, ground truth is taken from the detection exchange that contains the same robot utterance, as illustrated in Figure 1.

Performance is reported using two metrics: the Mean Absolute Error (MAE) and the macro-averaged F1 score, as in [11]. The MAE, with 0 as the minimal and optimal score, reflects how far-removed the predictions are from the ground truth. Because the dataset is unbalanced (40% of exchanges were rated as '3' and only 5% as '1'), we also report the macro-averaged F1 score—the average of the F1 scores calculated separately for each of the five points of the scale. This score ranges between 0 and 1, with 1 being the optimal score, and penalises more heavily a system that would always predict the most common value, mitigating the effect of the class imbalance in the results. To show this difference, we also report the scores of a baseline system that always predicts the majority rating.

To prevent overfitting on this dataset, we split up the 25 interactions into a "development" set of 15 interactions (totalling 355 exchanges) and a "test" set of 10 (totalling 235 exchanges), following common machine learning practices. All evaluations of the prompt components are performed on the development set, while the main two conditions (prediction against detection) are evaluated on the test set.

C. System

The enjoyment predictions were generated by OpenAI's LLM GPT-4o, specifically using the `gpt-4o-2024-08-06` checkpoint, as this is generally recognised as the LLM with the best text comprehension at the time of writing [19]. In the *detection* condition (including the user response), the model was prompted using the following text, as adapted from [11]:

> *Given the following scale and the current exchange between a robot and a human, rate the user enjoyment in the current exchange with an integer value (1 to 5).*

For the *prediction* condition (omitting the user response), the prompt was adapted by replacing *"rate the user enjoyment in the current exchange"* with *"predict the user enjoyment after this exchange"*.

This prompt was followed by the description of the scale given in Section III-A. In an ablation study, we evaluate the impact of the following additional input components, adapted from [11], on system performance:

- **Scale details:** a description of cues an annotator should look for in the dialogue, for each point of the scale.
- **Examples:** 10 example exchanges from different dialogues, each with an expert annotator's rating and reasoning for the chosen rating.
- **Reasoning:** prompting the model to output a reasoning for the chosen rating.
- **Dialogue context:** all previous exchanges in the dialogue.
- **Rating history:** for each previous exchange in the dialogue, the enjoyment rating that was predicted by the system.

The last element of the prompt is the exchange that is to be rated. The format of the exchange depends on the condition:

- **Detection (with user response):** robot utterance and subsequent user utterance.
- **Prediction (without user response):** user utterance and subsequent robot utterance.

IV. Results

First, we evaluate the contribution of each prompt component to the model's performance, evaluating turn-by-turn enjoyment prediction condition using the development set. Then, we compare the prediction condition with the detection condition to answer our main research question, evaluating turn-by-turn enjoyment on the test set. Finally, we evaluate the correlation with whole-dialogue user perceptions.

A. Ablation study

Table I reports the performance of the system on the development set when expanding the prompt with different input components. Rows starting with "+" indicate the prompt contained all components of the rows above.

The results show that, although the differences in scores are small, each of the added components increases performance on both metrics, except adding the history of the system's ratings.

Using the system with all components except rating history, we evaluated the impact of the initial sentence of the prompt, and changed it from the prompt that was adapted to the prediction condition, to the prompt that was also used for the detection condition. Surprisingly, this increased performance

TABLE I
ABLATION STUDY OF PROMPT COMPONENTS ($n_{exchanges} = 355$)

Prompt components	MAE	F1 score
Only simple prompt, adapted for prediction	0.95	0.22
+ scale details	0.95	0.24
+ examples	0.93	0.26
+ reasoning	0.88	0.26
+ dialogue context	0.87	0.27
+ rating history	0.90	0.23
Detection prompt, all components exc. rating history	**0.83**	**0.31**
+ examples include user response	0.84	0.29

TABLE II
PREDICTION AGAINST DETECTION AND HUMAN ANNOTATORS, USING ANNOTATOR 3 AS GROUND TRUTH ($n_{exchanges} = 235$, $n_{dialogues} = 25$)

Model	MAE	F1	Satisfaction	Strangeness
Prediction	0.81	0.25	0.59 (*)	-0.54 (*)
Detection	0.83	0.24	0.55 (*)	-0.56 (*)
Annotator 1	0.69	0.32	0.01	-0.16
Annotator 2	0.60	0.40	-0.05	-0.01
Annotator 3	0.00	1.00	0.01	-0.29
Majority Baseline	0.76	0.11	—	—

in the prediction condition. However, changing the format of the example exchanges from "user utterance-robot response" to "robot utterance-user response" decreased performance.

B. Prediction vs. Detection

We report the performance of our system on the test set, comparing its performance in our two main conditions: with access to the user response (detection) and without (prediction). We use the best model configuration as found in the previous section: using all prompt components except for rating history and use the detection prompt in both conditions. These results are reported in Table II and compared against the human expert annotators and a baseline that always predicts the rating that is most common in the dataset.

These results show that the prediction system is able to achieve the same performance as the detection system. Furthermore, this performance is even very close to that of Annotator 1. Although we use Annotator 3's ratings as ground truth, the system was not specifically designed to approximate their ratings, as it was only shown ratings all three annotators agreed on. This shows that the system approximates the performance of a human expert annotator, even without being shown the user response.

C. Whole-dialogue User Perceptions

We investigate to what extent the systems' ratings correspond to the users' own perception of enjoyment. As only ratings for the whole conversation are available, we average the systems' turn-by-turn ratings and calculate the correlation with the users' self-reported ratings. Performance is reported on the entire dataset ($n = 25$), as using whole-dialogue ratings leaves too little data to split into a development and test set.

In Table II, we report the performance of the same system configuration evaluated in the previous section, comparing prediction and detection, and the correlation between the expert annotators' whole-conversation enjoyment rating. All scores are Spearman correlations, and those that reach statistical significance are marked with (*) for $0.01 < p < 0.05$. The Bonferroni correction for multiple testing was applied.

While none of the human expert annotators' ratings significantly correlated with the users' perceptions, in both the prediction and detection condition, the system is able to achieve a significant moderate correlation with the users' self-reported satisfaction and perception of strangeness. This indicates that the system is clearly able to extract features that are relevant for the user. In neither condition did the system significantly correlate with fun or interestingness.

V. DISCUSSION AND CONCLUSION

In this work, we set out to investigate whether it is possible to predict user enjoyment in human-robot dialogue turn-by-turn, without seeing the user's response to the robot utterance. While essential for adaptive conversational systems, this issue was not yet considered by prior research. We built a system using an LLM for this task, opting for OpenAI's GPT-4o, and investigated which input data this LLM needs to perform best at this task. We have publicly released our system.

Our results show that prediction is possible: the system performed just as well when predicting enjoyment as when doing detection, where it did have access to the user response, surpassing expectations. Comparing with multiple expert annotators, it even achieved similar performance to the humans.

We also investigated whether these predictions actually correlate with the user's perception of the conversation. While human expert ratings did not correlate with user perceptions at all, our system again surpassed our expectations and achieved a significant correlation with user satisfaction and strangeness.

Multimodal information (e.g., video, audio) might provide benefits to predicting user enjoyment, by giving more information about dialogue context. However, this is a complex endeavour, as user feedback is subtle and fast. Future work should investigate how multimodal systems can be used for enjoyment prediction, as well as looking at smaller, local models that can better fulfil privacy, energy, and latency requirements. More detailed analysis of which aspects of the dialogue the LLMs makes use of also remains to be done.

Rating human-robot dialogue on a scale for user enjoyment remains a complex and subjective matter, even with a rigorously developed and validated scale. Although the MAE indicates that, on average, the model's prediction is less than 1 point removed from the expert annotation, the F1 score remains relatively low. However, this matches human performance and the model also correlates significantly with user perceptions. This shows that LLMs have sufficient representational ability of dialogue and user enjoyment to enable judgments that can improve dialogue without seeing the user response, either during a conversation or by applying adaptation methods to an LLM. Together with our tool, we believe these results will fast-forward research in adapting LLMs for conversational human-robot interaction.

References

[1] B. Irfan, S. Kuoppamäki, and G. Skantze, "Recommendations for designing conversational companion robots with older adults through foundation models," *Frontiers in Robotics and AI*, vol. 11, p. 1363713, 2024.

[2] E. Verhelst, R. Janssens, T. Demeester, and T. Belpaeme, "Adaptive second language tutoring using generative ai and a social robot," in *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 1080–1084, 2024.

[3] B. Irfan, S.-M. Kuoppamäki, and G. Skantze, "Between reality and delusion: challenges of applying large language models to companion robots for open-domain dialogues with older adults," 2023.

[4] H. H. Clark, *Using language*. Cambridge university press, 1996.

[5] A. Axelsson and G. Skantze, "Do you follow? a fully automated system for adaptive robot presenters," in *Proceedings of the 2023 acm/ieee international conference on human-robot interaction*, pp. 102–111, 2023.

[6] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, *et al.*, "Training a helpful and harmless assistant with reinforcement learning from human feedback," *arXiv preprint arXiv:2204.05862*, 2022.

[7] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, *et al.*, "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, vol. 35, pp. 27730–27744, 2022.

[8] B. Irfan, M. Staffa, A. Bobu, and N. Churamani, "Lifelong learning and personalization in long-term human-robot interaction (leap-hri): Open-world learning," in *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 1323–1325, 2024.

[9] B. Irfan, J. Miniota, S. Thunberg, E. Lagerstedt, S. Kuoppamäki, G. Skantze, and A. Pereira, "Human-robot interaction conversational user enjoyment scale (hri cues)," *arXiv preprint arXiv:2405.01354*, 2024.

[10] J. Deriu, A. Rodrigo, A. Otegi, G. Echegoyen, S. Rosset, E. Agirre, and M. Cieliebak, "Survey on evaluation methods for dialogue systems," *Artificial Intelligence Review*, vol. 54, pp. 755–810, 2021.

[11] A. Pereira, L. Marcinek, J. Miniota, S. Thunberg, E. Lagerstedt, J. Gustafson, G. Skantze, and B. Irfan, "Multimodal user enjoyment detection in human-robot conversation: The power of large language models," in *Proceedings of the 26th International Conference on Multimodal Interaction*, ICMI '24, (New York, NY, USA), p. 469–478, Association for Computing Machinery, 2024.

[12] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, *et al.*, "Judging llm-as-a-judge with mt-bench and chatbot arena," *Advances in Neural Information Processing Systems*, vol. 36, pp. 46595–46623, 2023.

[13] M. Paetzel-Prüsmann, J. F. Lehman, C. J. Gomez, and J. Kennedy, "An automatic evaluation framework for social conversations with robots," in *Proceedings of the 2024 International Symposium on Technological Advances in Human-Robot Interaction*, pp. 56–64, 2024.

[14] M. A. M. Shaikh, H. Prendinger, and I. Mitsuru, "Assessing sentiment of text by semantic dependency and contextual valence analysis," in *Affective Computing and Intelligent Interaction: Second International Conference, ACII 2007 Lisbon, Portugal, September 12-14, 2007 Proceedings 2*, pp. 191–202, Springer, 2007.

[15] J. Wang, L.-C. Yu, K. R. Lai, and X.-j. Zhang, "A locally weighted method to improve linear regression for lexical-based valence-arousal prediction," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 415–420, IEEE, 2015.

[16] B. Han, C. Yau, S. Lei, and J. Gratch, "Knowledge-based emotion recognition using large language models," *arXiv preprint arXiv:2408.04123*, 2024.

[17] Z. Liu, K. Yang, Q. Xie, T. Zhang, and S. Ananiadou, "Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5487–5496, 2024.

[18] A. N. Tak and J. Gratch, "Gpt-4 emulates average-human emotional cognition from a third-person perspective," *arXiv preprint arXiv:2408.13718*, 2024.

[19] OpenAI, "Gpt-4 technical report," 2023.